Deep Reinforcement Learning-based Policy for Baseband Function Placement and Routing of RAN in 5G and Beyond

Zhengguang Gao, Shuangyi Yan, Jiawei Zhang, Bingtao Han, Yongcheng Wang,

Yuming Xiao, Dimitra Simeonidou, Yuefeng Ji

***Abstract*—In this paper, we propose a deep reinforcement learning (DRL)-based algorithm to generate policies of Baseband Function (BBF) placement and routing. In order to explore the performance of the proposed algorithm in practical systems, the online scenario with the completely random requests is used in the simulation considering C-RAN and NG-RAN architectures. Besides, an Integer Linear Programming (ILP) model is formulated to generate the optimal solution as the benchmark. The simulation results show that DRL-based algorithm converges in a short time, and its performance closes to the optimal benchmark obtained by ILP in terms of latency and bandwidth for the online scenarios. In addition, the performance of the generated policies based on DRL is compared with a classic heuristic algorithm, i.e., first-fit algorithm. The performance of DRL-based algorithm is superior to the first-fit algorithm from above two perspectives. The fast convergence and the near-optimal performance prove that the DRL-based algorithm is a promising approach for the BBF placement and routing of RAN in 5G and Beyond.**

***Index Terms*— Deep reinforcement learning, Baseband Function placement and routing, 5G and Beyond.**

I. INTRODUCTION

THE design of 5G network is expected to provide diversified services including enhanced Mobile Broadband (eMBB), ultra-Reliable & Low Latency Communication (uRLLC), and Massive Machine Type Communication (mMTC) [2]. These services make 5G network radical changes in terms of stringent requirements of bandwidth, latency, and networking flexibility [3], [4]. Centralized radio access network (C-RAN) is a promising network architecture, which allows Baseband Units (BBUs) centralized in central offices (COs) [5]. With this design, baseband resource can be shared among several base stations, and OpEx can be significantly reduced through the centralized management and maintenance. Despite the advantages of C-RAN, it raises some challenges. The C-RAN's fronthaul which transmits raw I/Q sample data between BBUs and remote radio units (RRUs) through the fixed point-to-point connection suffers from high pressure to provide huge bandwidth, especially for future 5G communications. In addition, fronthaul supported by the load-independent common public radio interface (CPRI) limits the scalability and bandwidth efficiency of C-RAN. Therefore, eCPRI is proposed as a data-dependent fronthaul interface for data transmission on a frame basis, e.g., with Ethernet protocol, which remarkably economizes the fronthaul bandwidth than constant-bit based CPRI.

In order to support a high-bandwidth transmission and improve the scalability of RAN in 5G/B5G, the fronthaul of next-generation (NG-RAN) evolves from the "point-to-point" connection to the "any-to-any" connection [6], [7]. The high flexibility not only saves the fronthaul bandwidth significantly, in particular to massive MIMO scenarios [8], but also enables RRUs to share the computational resources in different COs. In addition, the flexibility of mobile fronthaul is more conducive to the deployment of advanced technologies such as coordinated multipoint (CoMP) transmission/reception, enhanced Inter-Cell Interference Coordination (eICIC), and BBU aggregation [9], [10]. More importantly, with the emerging advanced techniques introduced in the network such as mobile edge computing (MEC) and network function virtualization (NFV), network automation and intelligence are the necessary elements for 5G/B5G communications to improve the efficiency of network management and maintenance [11], [12].

Due to the evolution of RAN, it is necessary to adjust the corresponding policy of BBF placement and routing which decides the appropriate positions of BBFs and the lightpath provisioning from the RRU to data centre (DC). The adjusted policy should incorporate harmoniously with the evolution of

Parts of this work appeared in the proceeding of OFC 2019, San Diego, United State [1].

This work was supported by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (Grant No. 62021005). The Beijing Natural Science Foundation (No.4192039). It was also supported by EU H2020 grant (5G-Clarity, No. 871428).

Zhengguang Gao, Bingtao Han, Yongcheng Wang are with the State Key Laboratory of mobile Network and Mobile Multimedia Technology, 518055, China.

Zhengguang Gao, Jiawei Zhang, Yuming Xiao, and Yuefeng Ji are with the State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, 100876, China (corresponding author: Jiawei Zhang, email: zjw@bupt.edu.cn).

Shuangyi Yan and Dimitra Simeonidou are with the High Performance Networks Group, Smart Internet Lab, University of Bristol, Bristol, BS8 1TH, U.K.

RAN. On the one hand, traditional heuristic algorithms are difficult to achieve an ideal performance under dynamic network scenarios. The heuristic-based policy which adopts predefined procedures tends to stop searching for better policies once it gets an available solution [13]. On the other hand, ILP models can be formulated for a set of requests to search for the global optimal solution, however, the required time and computational resource of the ILP method prohibit the real-time deployment in practical networks. In addition, the ILP model uses a pre-set request matrix in an offline scenario, whereas the decision-making policy should run in an online scenario to handle the unpredictable changes of network state in practical networks. Therefore, it is necessary to introduce a more intelligent policy for BBF placement and routing, which abstracts the network state and adjusts the BBF placement and routing automatically. The proper algorithms should be able to achieve self-optimization and iterative upgradation. Motivated by the above, we tend to propose a flexible and self-learning BBF placement and routing policy based on DRL. As far as I know, this is the first work to explore the DRL-based policy for BBF placement and routing in an online scenario with complete random requests.

We extend our previous work about the DRL-based algorithm for BBF placement and routing in C-RAN with offline scenarios, where the request matrix is predefined and fixed [1]. To keep pace with the evolution of RAN and explore the intelligent policy, we implement a DRL-based algorithm for BBF placement and routing to optimize network resource utilization and transport latency in an online scenario both for C-RAN and NG-RAN. Besides, the proposed policy is formulated as an ILP model in detail. The result of ILP is used as an optimal benchmark for the proposed DRL-based policy. The proposed DRL-based algorithm is also compared with the first-fit algorithm. We analyse the performance of these three algorithms in terms of bandwidth, transport latency and BBF aggregation gain. The result shows that the DRL-based algorithm not only converges quickly in an online scenario, but also achieves the satisfactory performance.

The remainder of this paper is organized as follows. In Section II, we present some related works and discuss the background of our work. The system model is described in Section III. The ILP model and DRL-based algorithm are presented in Section IV. Section V gives a detailed description of simulation results. We make a conclusion in Section VI.

II. Related Works

The problem of BBF placement in RAN has attracted considerable research interests in the literature. A fixed/mobile convergence aggregation optical network was proposed in [14]. The authors formulated an optimization problem to minimize the aggregation infrastructure power. Reference [15] presented the problem of BBU placement over a wavelength division multiplexing (WDM) aggregation optical network. The author formulated an ILP model to evaluate two different fronthaul transport cases. These two works optimized BBU placement in WDM optical networks. Reference [16] proposed to place BBU pools at the edge of the network. The authors solved the BBU placement problem over the wireless front–hauls. An ILP-based algorithm and a heuristic algorithm were developed for small and large networks respectively. Reference [17] proposed a Digital Unit pool placement problem. A Mixed Integer Linear Programming (MILP) model is formulated to minimize the total cost of ownership. These two works tried to solve the BBU placement problem through different concepts, wireless front-hauls in [16] and Digital Unit pool in [17]. The architecture of survivable RAN was widely discussed to solve the problem of link failures in [18]-[21]. Reference [18] introduced an efficient and proactive restoration mechanism to ensure service resilience under the tremendous mobile traffic in carrier clouds. The authors in [19] used ILP and Branch-and-Price algorithms to optimize virtualized BBU selection problems based on resiliency and price. Reference [20] proposed a survivable fronthaul scheme against single hotel failure. The authors used both ILP and heuristic methods to evaluate different strategies of survivable BBU placement. Reference [21] introduced three protection strategies: dedicated path protection, dedicated BBU protection, and dedicated path and BBU protection. The authors formulated an ILP model to minimize the consumption of computational resources, the number of used wavelengths and BBU pools. To overcome the unpredictable failures from the links and BBU hotels, various protection schemes were proposed in these works to ensure the resiliency of BBU placement.

With the evolution of RAN to address the rapid growth of fronthaul traffic and the stringent requirement of latency, more flexible RAN architectures such as NG-RAN have been analyzed. Reference [22] proposed a resource allocation policy for RAN slicing in Multi-CU/DU architectures, the result showed that the proposed method reached the satisfactory performance and guaranteed the isolation between slices. The effective management strategy for the agile DU-CU deployment was investigated in terms of power consumption in [23], a mixed ILP model and a graph-based heuristic method was proposed to optimize the consumption of reconfigurable add/drop multiplexer, Ethernet switch, optical transponder and so on. With the increasing complexity and dynamicity of network, it is difficult to find an adaptive and efficient strategy to optimize resource allocation for network.

Recently, deep reinforcement learning (DRL) has gained increasing attention after AlphaGo defeated the world's best chess player [24]. DRL that combines Deep Learning and Reinforcement Learning together can handle complicated problems because the former processes complex information and the latter optimizes complex decision-making strategies [25]. This promising methodological paradigm has been explored in resource allocation for 5G networks. The authors in [26] proposed a multi-agent DRL-based algorithm for service provisioning of multi-domain optical networks. The result showed that the proposed framework outperformed the existing rule-based heuristic algorithms significantly. A DRL-based algorithm was proposed to optimize network slicing problem in [27]. The results showed the proposed policy outperformed benchmark heuristics in terms of the profit of infrastructure providers. The authors in [28] designed a DRL-

based framework to address the problem of virtual network (VNT) slicing in datacenter interconnections. The experiment showed that the proposed framework can provision VNT requests with shorter time and comparable blocking performance. An online multi-tenant secret-key assignment policy based on DRL was proposed in [29]. The proposed method reduced the tenant-request blocking probability significantly. A DRL-based algorithm was proposed to accommodate diversified services in 5G/B5G networks in [30]. The results showed that the proposed algorithms outperformed the benchmark by ILP and widely used heuristics significantly considering the resource-saving and the service-scale.

DRL can perform end-to-end training and abstract a complex multi-layered model from the state to action. It is considered as one of the key-enabling technologies to solve sequential decision-making problems. In addition, the random process of request can be introduced in the training process of DRL-based algorithms. Therefore, the DRL-based method can be designed as an online algorithm to generate policy for BBU placement and routing of flexible RAN.

III. RAN ARCHITECTURE

Fig. 1 shows the typical architecture of RAN. A certain number of RRUs are aggregated to several COs, and each CO is interconnected by optical links, which constitute a converged wireless and optical aggregation network. In 5G, eCPRI is considered as an important interface protocol to support much more fronthaul traffic. It enables more flexibility of the functional split for the physical layer, and puts some low physical functions into RRU. Generally, the data of a request from RRU is sent to the CO with the eCPRI encapsulation to do the baseband processing, then the processed data is transported to data center (DC) for core network processing. Therefore, a general service should include the complete BBF processing and the routing from RRU to DC.

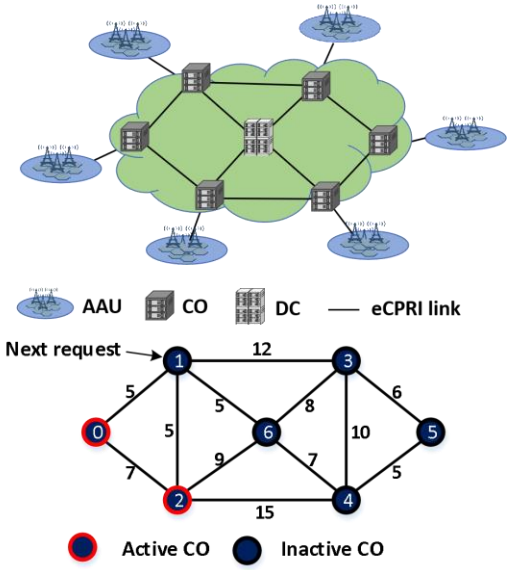

Fig. 1. Architecture of RAN.

C-RAN was gradually adopted in 4G+ and the initial stage of 5G due to its advantages in energy consumption, operation and maintenance costs, and it will be massively deployed in the mature stage of 5G era. To address the issue of sharp increase in fronthaul traffic, 3GPP has also advocated new standards for NG-RAN [23], [31]. NG-RAN disaggregates partial physical-layer BBFs from BBU to the cell site to reduce the fronthaul bandwidth consumption, while dividing the rest of BBFs into a distributed unit (DU) and a central unit (CU). This design can also reduce the fronthaul latency since DU is placed closer to RRU than traditional BBU, while guaranteeing the centralization gain through CU aggregation in remote COs. The disaggregation of BBF has improved the RAN flexibility for diversified service demands, but also increased the difficulty of BBF placement and routing. Thus, we consider that it is necessary to explore more effective and adaptive policy to plan the RAN in 5G/B5G, especially based on many advanced technologies and concepts that may bring the technological revolution like artificial intelligence. Here we try to formulate a unified ILP model following these points: 1) Active COs which have been switched on are chosen preferably to hold the BBF for high consolidation. 2) The CO closest to the requests is chosen to host BBFs. This scenario benefits to the reduction of fronthaul latency and the cost of mobile fronthaul bandwidth. 3) We choose the shortest path from RRU to DC for each request. In some cases, all these points can't be fulfilled together. For example, node 0 and node 2 are active in Fig. 1. The next request can't select the closest CO, i.e., node 1, as it is inactive. The active COs are also not on the optimal routing path from node 1 to node 6. Therefore, we need to cope with the trade-off between the above three points and search the best solution for the request sequence. In the next section, we developed an ILP model to search the optimal solution for BBU placement and routing in an offline scenario, the unified ILP model can consider both C-RAN and NG-RAN through adapting the pattern of RAN slicing. We then explored the DRL-based algorithm for tacking this complicated decision-making problem in an online scenario.

IV. ILP MODEL AND DRL-BASED ALGORITHMS

A. The Description of ILP Model

In this section, ILP model for flexible BBU placement and routing is developed in detail. In order to generate the solution fully consistent with the ideas mentioned above, we establish a multi-objective optimization function to minimize the consumption of COs, bandwidth and transport latency. And the result of ILP is also used as an optimal benchmark for DRL-based algorithm.

(A) Inputs and parameters

a) B: Set of services.
b) R: Set of COs, which can hold BBU.
c) D: Data centre.
d) E: Set of optical links.
e) P: Set of paths from RRU to DC.
f) K: Number of BBU Functions. ($K = 3$ represents

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2021.3110788, Journal of Lightwave Technology

the C-RAN, $K=4$ represents NG-RAN. The first BBF is in RRU i.e., the origin of services, the last BBF equals to the content processing in destination DC).

g) T_r : The computational resource (GOPS, Giga operations per second) of CO $r \in R$.
h) C_e: The bandwidth (Gbps) of optical links $e \in$ E.
i) T_e: The transport latency of optical links $e \in$ E.
j) T_b: The maximum latency allowed for the service $b \in$ B.
k) $B_{b,k}$: The cost of bandwidth (Gbps) of the service $b \in$ B after the k-th BBF processing.
l) $C_{b,k}$: The requirement of computational resource (GOPS) for the k-th BBF processing of service $b \in$ B.
m) $G_{p,e}$: Binary indicator, 1 if the path $p \in$ P includes the link e.
n) $M_{r,e}$: Binary indicator, 1 if the link $e \in E$ is connected to CO r.
o) $Base_{b,r}$: Binary indicator, 1 if the RRU with current service $b \in$ B is connected to CO $r \in R$.
p) L_p: The number of links on path $p \in$ P.
q) $PN_{p,r}$: Binary indicator, 1 if the path $p \in$ P passes the CO $r \in R$.

(B) Variables of model

a) $U_{b,p}$: Binary variable, 1 if service $b \in$ B selects the path $p \in$ P.
b) $Z_{b,k}^{e,p}$: Binary variable, 1 if service $b \in$ B is carried on the link $e \in E$ of path $p \in$ P after the k-th BBF processing.
c) $O_{b,k}^{r}$: Binary variable, 1 if the k-th BBF of service $b \in$ B is processed in CO $r \in R$.
d) B_r : Binary variable, 1 if CO $r \in R$ is active.

(C) Objective function: To consider the mentioned concepts, a multi-objective optimization function is formulated as:

$$min\ (\alpha \times \sum_r B_r + \beta \times \sum_{b,k,e,p} Z_{b,k}^{e,p} \times B_{b,k} + \gamma \times \sum_{b,k,e,p} Z_{b,k}^{e,p} \times T_e). \quad (1)$$

The objective function consists of three parts. The first term is to minimize the number of active COs. The second term is to minimize the cost of bandwidth on all links. The third part aims at optimizing the transport latency of all services. We can change the priority of these three factors in the optimization process by adjusting the weights (α, β, γ).

(D) Constraints:

- Routing:

$$\sum_p U_{b,p} = 1, \forall\ b \in \text{B}. \quad (2)$$

$$\sum_{k,e,p} Z_{b,k}^{e,p} = \sum_p U_{b,p} \cdot L_p, \forall b \in B. \quad (3)$$

- Capacity:

$$\sum_{b,p} \sum_{k \neq K} Z_{b,k}^{e,p} \times B_{b,k} \leq C_e, \forall e \in \text{E}. \quad (4)$$

$$\sum_b \sum_{k \neq K} O_{b,k}^{r} \times C_{b,k} \leq T_r, \forall r \neq DC. \quad (5)$$

- BBU placement:

$$\sum_r O_{b,k}^{r} = 1, \forall\ b \in \text{B}, \forall k \in K. \quad (6)$$

$$O_{b,k}^{r} = \begin{cases} Base_{b,r}, \text{if } k = 1 \\ 1, \text{if } k = K \end{cases}, \forall\ b \in B. \quad (7)$$

$$O_{b,k}^{r} = \sum_{k \leq k1 \leq K} Z_{b,k1}^{e,p}, if\ Base_{b,r} = 1, \forall b \in B, r \in R, k \in K. \quad (8)$$

$$2 \times O_{b,k}^{r} \leq \sum_{e,p} \sum_{k1=1}^{k1=k-1} Z_{b,k1}^{e,p} \times M_{r,e} + \sum_{e,p} \sum_{k2=k}^{k2=K-1} Z_{b,k2}^{e,p} \times M_{r,e} \leq O_{b,k}^{r} + 1, \text{if } Base_{b,r} = 0\ and\ PN_{p,r} = 1, \forall b \in \text{B}, p \in \text{P}, r \neq DC, k \neq K. \quad (9)$$

$$B_r \leq \sum_b \sum_{k1=2}^{k1=K-1} O_{b,k1}^{r} \leq Max \cdot B_r, \forall r \in R. \quad (10)$$

- Latency:

$$\sum_{e,p} \sum_{k1=1}^{k1=K-2} Z_{b,k1}^{e,p} \times T_e \leq T_b, \forall b \in \text{B}. \quad (11)$$

Equation 2 ensures that service $b \in$ B selects only one path to DC. Equation 3 ensures that service traffic passes through all the links of selected path $p \in$ P. Equation 4 guarantees that the bandwidth requirement of all services carried on each link does not exceed its capacity. Equation 5 ensures that the cost of computational resource for processing BBFs of all services does not exceed the capacity of each CO. Equation 6 guarantees that all the BBFs of service $b \in$ B have been processed. Equation 7 ensures the first and the last BBF of service $b \in$ B is processed in its BS and DC respectively. Equation 8 and 9 decide the position of BBFs processing for service $b \in$ B. The position depends on the state of ingress and egress flows. Equation 10 identities that once the k-th $(k \neq K, k \neq 1)$ BBF of service $b \in$ B is processed in CO $r \in$ R, then $B_r = 1$. Equation 11 ensures the fronthaul latency of service $b \in$ B does not exceed its threshold.

B. DRL-based Algorithm

The structure of DRL-based policy is shown in Fig. 2. Deep double Q-Learning algorithm [32] is deployed to fit the Q-value $Q(s_t, a_t)$ which is a measure of the overall expected reward when the agent in state s_t performs action a_t during a completed episode. In the experiment, the value of $Q(s_t, a_t)$ can reflect whether the action a_t composed of BBU placement and routing path is suitable for the current network state s_t represented by the bandwidth of optical links, the

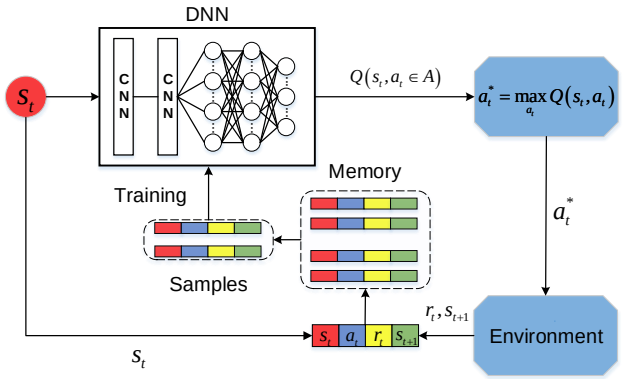

Fig. 2. Structure of the proposed DRL-based algorithm.

computational resource of each CO in RAN, and the request position. Once the accurate Q-function $Q(s_t, a_t)$ is formulated by DNN with parameters $\boldsymbol{\theta}$, we can find the best mapping from the current network state s_t to the corresponding action by $a_t^* = \max_{a_t} Q(s_t, a_t)$. The DRL-based algorithm has the following three steps.

The first step is to create the environment which can simulate the scenario of the actual network. The environment must reflect the accurate variations of bandwidth, computational resource from the current state s_t to next state s_{t+1} and the instant return $r(s_t, a_t)$.In order to make decision for BBF placement and routing in RAN, the action a_t should contain the node information for BBF processing and routing information. Generally, a vector v containing N elements can represent a certain action as $v = [v_1, v_2, ..., v_N]$. The first element v_1 denotes the location of BBU in C-RAN, the other elements v_2 to v_N can decide the routing from one of RRUs to DC. The last element of v_N must be DC, so it exists C_{n-1}^{N-1} possible combinations for other $N-1$ elements from the nodes except DC, where n is the number of nodes. Therefore, the upper bound for the cardinality of the action space can be estimated as n^{N-1} through some simple mathematical derivation. The difference between C-RAN and NG-RAN is that the first two elements of sequences rather than one element are required to decide the locations of CU and DU. It means that it may need one more element to represent the action. And the upper bound of action's space for NG-RAN can be estimated as n^N.

The action can be arranged generally as it is described above, however, the action space increases with the number of nodes rapidly according to the upper bound for the cardinality of the action space, and it also contains amounts of invalid actions with inexistent paths and impossible placement for BBF processing. The large space tends to decrease the speed of algorithm convergence. Therefore, the action space is reduced significantly through selecting the *k* shortest paths in the experiment. For a certain request from any BS, we use the **"**k-shortest-paths**"** algorithm [33] to select the *k* shortest paths from RRU to the data center (DC), where $k = 3$ in this paper. Therefore, the actions for the request from RRU include one path from the path set and the node information for BBF processing on the selected path. All the information for the action is still represented as a sequence, but the action space only contains the possible action which includes an existed path and the corresponding node in this path.

The state which contains the information of network is easier to be represented. A two-dimensional matrix is used to represent the available bandwidth for the links of the state s_t. For the topology of 7 nodes in Fig. 1, a 7 by 7 matrix is enough to express the available bandwidth. If a physical link exists between these two nodes, the corresponding element equals to the available bandwidth of the link. Otherwise, it is set to 0. For the computational resource, a one-dimensional matrix is used to represent the available resource for each node. Finally, we use a one-dimensional matrix to represent the request position through one-hot encoding method. The request position is to denote the source CO of each request. The element with the position of a request in the topology is set to 1, the other elements are set to 0. Therefore, the state is comprised of the matrices for the available bandwidth, available computational resources and the request position. In addition to the state information, the connection information of the network topology is used as an initial information to establish the environment for DRL simulation, which decides the latency of the selected path. The $r(s_t, a_t)$ is defined as:

$$r(s_t, a_t) = \begin{cases} -(\alpha \times x_t + \beta \times b_t + \gamma \times l_t), & a_t \text{ is accepted} \\ R, & a_t \text{ is not accepted} \end{cases}.$$

Where the x_t is the binary indicator, $x_t = 1$ means that one inactive CO is used to hold BBU while $x_t = 0$ means that the CO selected to hold BBU has been activated. b_t is the cost of bandwidth for the current service. l_t is the transport latency of routing path. The parameters (α, β, γ) are the weights for the multi-objective function in ILP model. When the selected action a_t cannot meet the requirements of the current request, the instant reward is set to R, where R should be obviously lower than the instant rewards of the action that can be accepted. In our simulation, R are set to -200 and -500 respectively for C-RAN and NG-RAN. This design helps the algorithm to converge effectively.

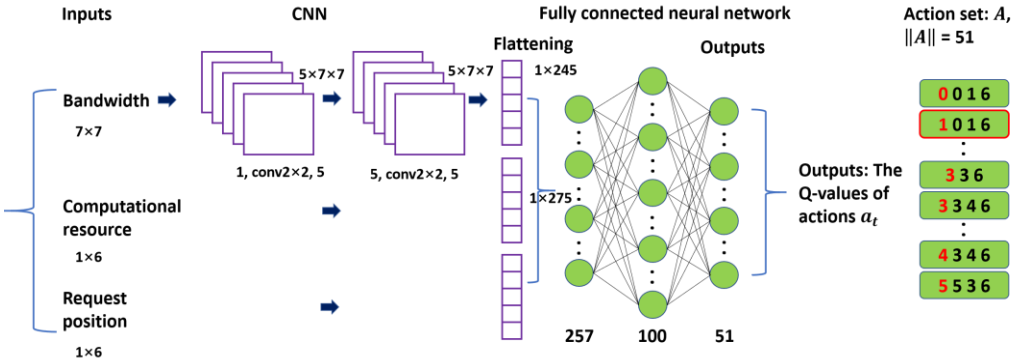

Fig. 3. Framework of Deep learning in DRL-based algorithm.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2021.3110788, Journal of Lightwave Technology

TABLE I
THE PSEUDOCODE OF PROPOSED DRL-BASED ALGORITHM.

Algorithm: DRL-based algorithm for BBU placement and routing

Input: Set of links = {start node, end node, link length, bandwidth capacity}; COs = {capacity of GOPS for each CO}; Requests = {RRU's position, GOPS demand, and bandwidth demand of fronthaul and backhaul}.

Output: the action a_t^* for BBU placement and routing for current state s_t .

Initial the parameters of DNN with Gaussian distribution and the memory with 3000 spaces.
Set episode number M and the maximum pace P allowed for one episode.
for _ in 1, 2, 3, …, M:
  Get the initial network state s_t and initial step p_t from the reset of environment.
  Initial the request b_t
  While $p_t \leq P$:
    DNN outputs the Q-value $Q^{\pi_t}(s_t, a_t, \boldsymbol{\theta_t})$ under the current s_t and request b_t.
    Choose the action a_t by the probability ε with maximum Q-values and the probability $1-\varepsilon$ with random action from action space A.
    Use the action a_t to interact with environment, get the instant return $r(s_t, a_t)$ and s_{t+1}.
    Send the data $[s_t, a_{t,} r(s_t, a_t), s_{t+1}]$ to the memory.
    If the number of data N > 3000:
      Update DNN parameters by stochastic gradient descent: $\boldsymbol{\theta_{t+1}} \leftarrow \boldsymbol{\theta_t} - \mu \sum \boldsymbol{d}\, \xi^i / d\boldsymbol{\theta_t}$ based on a batch of data sampling from the memory.
    $s_t = s_{t+1}$.
    $\varepsilon = \varepsilon + 0.000003$ (greedy_increment).
    If the current request is accepted:
      The next request is generated;
      $p_t = p_t + 1$;
    else: the next request remains the same.
  End
End

The second step is to prepare the data $[s_t, a_{t,} r(s_t, a_t), s_{t+1}]$ by interacting with the environment. As in Fig. 2, the current network state s_t is sent to DNN, and DNN with parameters $\boldsymbol{\theta_t}$ outputs the Q-values of all the actions, $Q(s_t, a_t), a_t \in A$. The action a_t with the largest Q-value is selected with the probability of ε to interact with the environment, and the environment outputs the instant return $r(s_t, a_t)$ and next state s_{t+1}. Finally, $[s_t, a_{t,} r(s_t, a_t), s_{t+1}]$ is saved in the memory of database. To represent the network state s_t more precisely, the specific deep learning framework is designed in Fig. 3. We formulate the inputs of network state as the matrix, and 2×2 convolution kernels are used to detect the feature of nodes whether it exists direct optical links between two nodes.

The detailed framework of neural network is described in Fig. 3. We first use two convolution layers to process the feature of bandwidth with the size of 7×7 for the topology in C-RAN. Each convolution layer contains 5 channels with 2×2 kernel. The parameter of padding is set to be "same" for CNN layers, so the outputs of these two convolution layers are 5×7 ×7 respectively. Then the flattened output from CNN (1×245), the input of computational resource (1×6) and the request position (1×6) are combined as the feature with the size of 1× 257. After that, we build one fully connected layer with 100 hidden neurons. The tanh activation function is used for the hidden layers. Finally, we build the last layer containing 51 neurons with no activation function to estimate the Q-value of actions. The data structure of action is also showed in Fig. 3. In C-RAN, the position of BBU is decided by the first element of a_t, the routing information is stored in the rest elements of a_t. For example, the action [1 0 1 6] means that the current policy of routing is from node 0 to node 1, node 1 to node 6. The node 1 is selected as CO to hold the BBU. It should be noted that the only difference for the action in NG-RAN is that the action contains two nodes for the placement of DU/CU.

The third step is to train DNN until it converges. The principle for the training of DNN is the Bellman optimality equation as [34]

$$Q^{\pi_t}(s_t, a_t) = r(s_t, a_t) + \gamma \times max Q^{\pi_t}(s_{t+1}, a_{t+1}). \quad (12)$$

Where γ is a discount factor, which determines the importance of the return in the future. The Q-function fitted by neural networks with parameters $\boldsymbol{\theta_t}$ is used to estimate the value of Q-function as $Q^{\pi_t}(s_t, a_t) \approx Q^{\pi_t}(s_t, a_t, \boldsymbol{\theta_t})$, $Q^{\pi_t}(s_{t+1}, a_{t+1}) \approx Q^{\pi_t}(s_{t+1}, a_{t+1}, \boldsymbol{\theta_t})$, so the optimal fitting of Q-function by DNN should also satisfy the principle of Bellman optimality equation as

$$Q^{\pi_t}(s_t, a_t, \boldsymbol{\theta_t}) = r(s_t, a_t) + \gamma \times max Q^{\pi_t}(s_{t+1}, a_{t+1}, \boldsymbol{\theta_t}). \quad (13)$$

Here the Bellman error is defined as

$$\xi = Q^{\pi_t}(s_t, a_t, \boldsymbol{\theta_t}) - r(s_t, a_t) - \gamma \times max Q^{\pi_t}(s_{t+1}, a_{t+1}, \boldsymbol{\theta_t}). \quad (14)$$

Therefore, the parameters of DNN can be updated by minimizing the Bellman error through stochastic gradient descent techniques. The pseudocode of the proposed DRL-based algorithm is provided in TABLE I. It should be noted that the bandwidth of links is set to be enough in the online scenario. Therefore, the maximum number of requests N that the system can accommodate mainly depands on the cost of computational resource for a request and the total computational resource. In the simulation, N is estimated in advance, then we set the maximum pace P based on N, which satisfies this inequality $P \leq N$. Considering the extreme case that most requests origin from one single BS tends to exhaust the local resources of network along the corresponding BS, P is set to 96 in our simulation, which is lower than N.

V. SIMULATION AND DISCUSSION

The simulation of the DRL-based algorithm is demonstrated in this section. In order to analyze the unified performance of DRL-based algorithms, the algorithm is analyzed in two scenarios, C-RAN and NG-RAN. The topology of C-RAN is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2021.3110788, Journal of Lightwave Technology

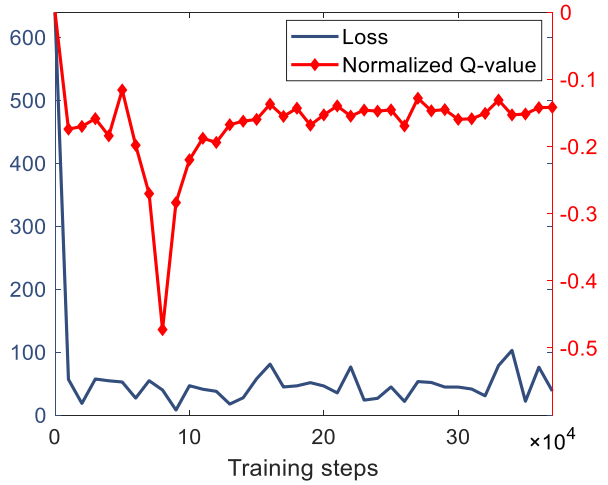

Fig. 4. Convergence of proposed DRL-based algorithm.

showed in Fig. 1, which contains 7 nodes. And the scenario of NG-RAN is more complicated, which contains 17 nodes. The bandwidth of each link is initialized to 25 Gbps for C-RAN, and the computational resource in each CO is initialized to 40000 GOPS. The bandwidth requirement of fronthaul and backhaul for a given request is 2.2948 Gbps and 0.2432 Gbps respectively with the wireless parameters of 100MHz spectrum, 2×2 MIMO and 16QAM for upstream. The cost of computational resource for BBFs of a request is 1200 GPOS. Based on the cost of computational resource for a request and the total computational resources, the maximum number of requests *N* that the system can accommodate is estimated as 192. Considering the extreme case that the most requests of an episode from a single RRU may exhaust the bandwidth of a link around this base station, the maximum pace *P* is set to 96. The maximum fronthaul latency allowed is 50 μs, which equals to 10 km in this paper. The weights (α, β, γ) for the ILP model and the instant return are set as $(100, 10, 1)$.

A. The simulation for C-RAN

In order to match the practical scenarios, the result of DRL-based algorithm is presented in an online scenario, where the request matrix is completely random that the next request

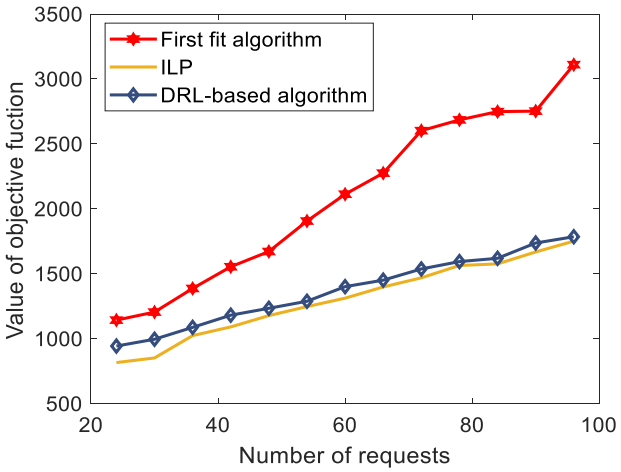

Fig.5. Value of objective function for different algorithms in C-RAN

origins from the arbitrary CO. We do the simulations of DRL-based and first-fit algorithms. Then the random request sequence is saved as the input of ILP model. In Fig. 4, the convergence presented by the loss of Bellman error and the normalized Q-value is plotted against the training steps. We can see that the loss of DNN drops quickly as the training begins, then converges a very small value. And the normalized Q-value decreases at first for the random exploration, then increases quickly for the improving policy by the training of DNN, finally reaches a large value which means that it has achieved a satisfactory policy. According to the definition of the instant return value $r(s_t, a_t)$, Q value is probably a large negative value for the currently random policy, so the normalized Q-value decreases at first for the random exploration, then increases quickly for the improving policy by the training of DNN, finally it reaches a large and stable value which means that it has achieved a satisfactory policy. It exists a small fluctuation both in the error curve and the normalized Q-value curve even though their values have converged. This is because we have introduced random exploration in DRL-based algorithms.

The value of objection function (1) which represents the weighted sum of bandwidth, the active CO's number and the transport latency is showed in Fig. 5. For the random requests

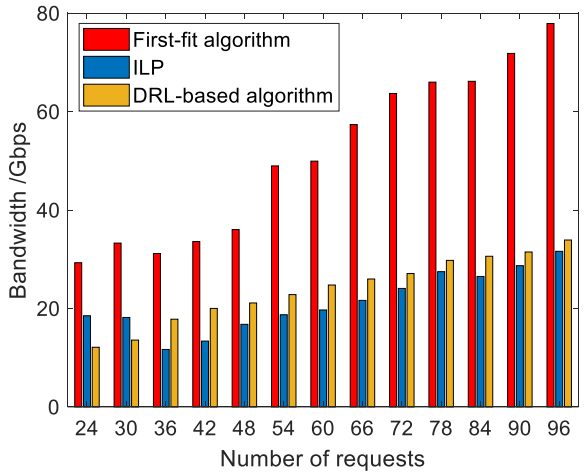

Fig. 6. Required bandwidth for different algorithms.

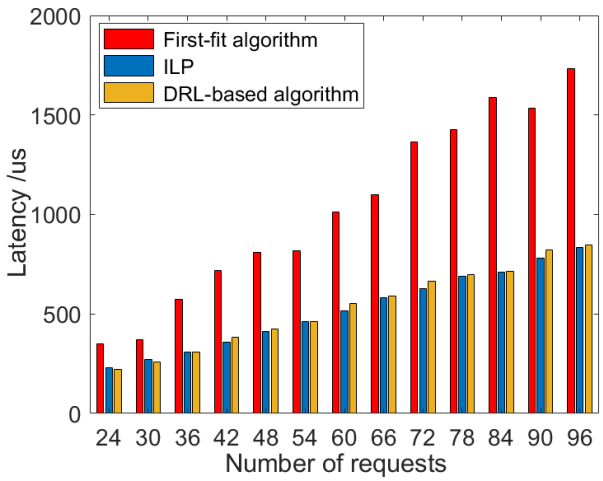

Fig. 7. Transport latency of different algorithms.

from 24 to 96, we find that the function value of DRL-based algorithms is almost the same as the ILP model, which are both much lower than that of the first-fit algorithm. This result proves that the policy of DRL-based algorithm nearly achieves the same performance of optimal benchmark by the ILP model. Based on Fig. 4 and 5, we can conclude that the proposed DRL-based algorithm not only converges quickly, but also achieves the sub-optimal results.

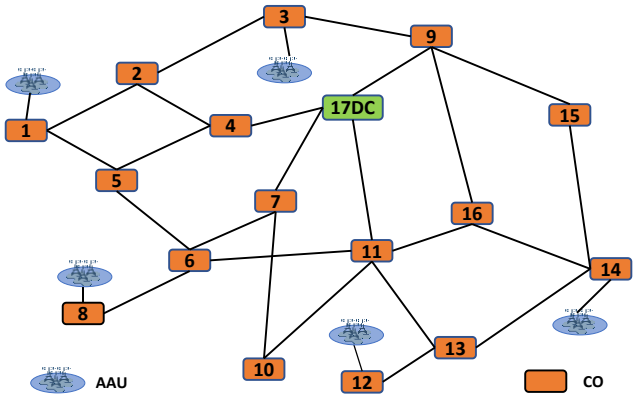

Fig. 8. The topology of NG-RAN.

In Fig. 6 and 7, the comparisons of cost for bandwidth and the transport latency are made between the first-fit algorithm, DRL-based algorithm and the ILP model. The figures show that the ILP model achieves the least cost of bandwidth and transport latency for the most cases. The cost for DRL-based algorithm is very close to the ILP model, and both of their cost are far less than the first-fit algorithm. In some scenarios like 24 and 30 requests, the cost of bandwidth and transport latency for DRL-based algorithm is less than the ILP model, the optimal benchmark. This is because the weights for the objective function of ILP model in (1) are set as $(100, 10, 1)$. It means that the most important factor for ILP model is the number of active COs (i.e., CO dominates the cost of whole network). Therefore, it firstly tends to uses fewer COs to reduce the cost of network. It is observed that the solution of the ILP model just uses 4 COs while the DRL-based algorithm occupies 6 COs for 24 and 30 requests. The ILP model searches the best solution to optimize the objective function while the DRL-based algorithm explores the possible strategy based the Q-value, which is deeply influenced by the instant return $r(s_t, a_t)$. According to the definition of the objective function and the instant return value, the strategies of ILP and DRL-based algorithms both can be adjusted by changing their weights.

B. The simulation for NG-RAN

In the NG-RAN scenario, the algorithm is required to the select the nodes for CU and DU. A more complex topology for the simulation is used in Fig. 8. Services can origin from 5 possible AAUs randomly. The bandwidth requirement of fronthaul, midhaul, and backhaul for a given request is 4.571 Gbps, 0.487 Gbps and 0.2432 Gbps respectively with the wireless parameters of 100MHz spectrum, 4×4 MIMO and 16QAM for upstream. The cost of computational resource for DU and CU of a request is 4600 GPOS and 200 GOPS. The maximum pace P is set to 96. The maximum latency allowed for fronthaul and midhaul is 100 μs and 500 μs, which equals to 20 km and 100 km in this topology [35]. The weights (α, β, γ) for ILP and DRL-based algorithm are consistent with the settings in C-RAN. Since the topology changes complicatedly with 17 nodes, the possible actions in DRL-based algorithms increase to 88 from 51 in C-RAN.

In Fig. 9 and 10, the comparisons of cost for bandwidth and the transport latency are made between first-fit, DRL-based algorithms and the ILP model in NG-RAN. The figures show that the ILP model achieves the best performance, which requires the least bandwidth and transport latency for all the cases. On the other hand, the cost for DRL-based algorithm is very close to ILP model. And these two algorithms outperform the first-fit algorithm significantly. According to Fig. 9 and Fig. 10, the required bandwidth for DRL-based algorithm is completely the same as the ILP model, the transport latency for DRL-based algorithm is a bit larger than the ILP model. The results denote that the proposed algorithm generates the similar policy for the placement of CU/DU, and plans the sub-

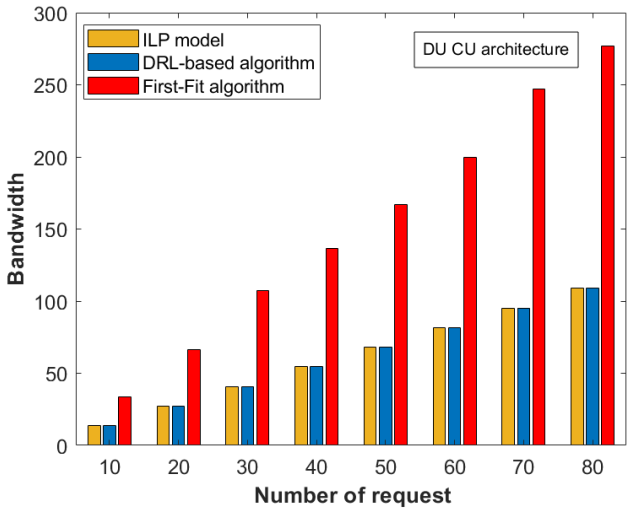

Fig.9. Required bandwidth for different algorithms in NG-RAN

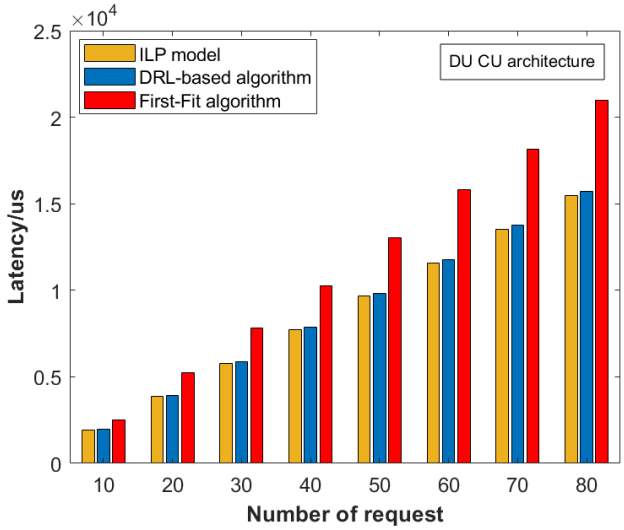

Fig.10. Required transport latency for different algorithms in NG-RAN

optimal routing for the system. The above results and analysis show that the DRL-based algorithm has potentials to address the decision-making problem in complicated scenarios.

C. Motivation of DRL-based algorithms in practical systems

According to the above results, it is concluded that the performance of DRL-based algorithm outperforms the first-fit algorithm significantly, and it reaches the sub-optimal performance compared with the ILP model. The ILP model can provide the optimal solution for a set of known requests, however, it is difficult to address the online scenarios with unpredictable requests because the request input for ILP must be predefined. On the other hand, the proposed DRL algorithm can adapt to the online scenarios. Through the introduction of the random requests in the environment of DRL algorithms, the data from the environment can instruct the agent to learn the corresponding policy for dynamic scenarios. This is the main motivation to explore DRL-based algorithms for BBU placement and routing of RAN, because there are unpredictable, dynamic and random events such as 4K live broadcast, VR/AR, telemedicine, which brings in dynamic traffic (time- and spatial-varying). And the final result in the paper has proved the proposed algorithm's effectiveness in an online scenario.

On the other hand, we should pay attention to the deployment of DRL-based algorithms in practical systems. The unavoidable problem of reinforcement learning is that it requires a certain amount of training time, and the time increases with the complexity of the problem. To train the DRL-based algorithms to generate the policy of BBF placement and routing in C-RAN and NG-RAN, it requires about 1 hour for the convergence of the algorithms with i5-8250U CPU and 8 GB RAM. The inference time for obtaining an action are microseconds, which can be ignored compared with the training time. Therefore, the main concern for deploying DRL-based algorithms is to reduce the training time. One possible solution is to prepare the pre-trained model in advance. In addition to this, the design of action space and the instant reward is very important for the fast convergence of DRL algorithms. The follow-up work will explore the influence of these factors on the algorithms.

VI. Conclusion

In this paper, we propose a DRL-based algorithm to solve the decision-making problem of BBU placement and routing for C-RAN and NG-RAN. In order to consider the possible dynamic events in practical systems, we design an online scenario with random requests to evaluate the proposed DRL-based algorithm. In addition, the first-fit algorithm and ILP model are also used as the baseline algorithms. The simulation results show that the proposed DRL-based algorithm converges effectively, and it outperforms the first-fit algorithm significantly both in offline and online scenarios. The DRL-based algorithm nearly reaches the optimal performance offered by ILP in terms of transport latency and bandwidth cost. In the future, the RAN network will be further evolved to more dynamic with integration of computing, perception, and communications. The dynamic traffic will cause more unpredictable events, which will require a self-learning policy to optimize the function placements. The performance of proposed DRL algorithm in an online scenario proves that DRL-based algorithms have potentials to address complicated decision-making problem. The proposed DRL algorithm are expected to provide the satisfactory performance for networks with dynamic events through implementing the training environment with random elements.

References

[1] Z. Gao, J. Zhang, S. Yan, Y. Xiao, D. Simeonidou and Y. Ji, "Deep Reinforcement Learning for BBU Placement and Routing in C-RAN," in 2019 Optical Fiber Communications Conference and Exhibition (OFC), San Diego, CA, USA, 2019, W2A.22.

[2] M. Agiwal, A. Roy and N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1617-1655, third quarter 2016.

[3] A. Tzanakaki et al., "Wireless-Optical Network Convergence: Enabling the 5G Architecture to Support Operational and End-User Services," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 184-192, Oct. 2017.

[4] Yuefeng Ji, Rentao Gu, Zeyuan Yang, Jin Li, Hui Li & Min Zhang, Artificial intelligence-driven autonomous optical networks: 3S architecture and key technologies , SCIENCE CHINA - INFORMATION, 2020, 63(6): 160301.

[5] J. Wu, Z. Zhang, Y. Hong and Y. Wen, "Cloud radio access network (C-RAN): a primer," *IEEE Network*, vol. 29, no. 1, pp. 35-41, Jan.-Feb. 2015.

[6] K. Tanaka and A. Agata, "Next-Generation Optical Access Networks for C-RAN," in Optical Fiber Communication Conference, OSA Technical Digest (online) (Optical Society of America, 2015), paper Tu2E.1.

[7] J. Zhang, Y. Ji, H. Yu, X. Huang, and H. Li, "Experimental demonstration of fronthaul flexibility for enhanced CoMP service in 5G radio and optical access networks," *Opt. Express*, vol. 25, no. 18, pp. 21247-21258, 2017.

[8] J. Zhang, Y. Xiao, D. Song, L. Bai and Y. Ji, "Joint Wavelength, Antenna, and Radio Resource Block Allocation for Massive MIMO Enabled Beamforming in a TWDM-PON Based Fronthaul," *Journal of Lightwave Technology*, vol. 37, no. 4, pp. 1396-1407, Feb. 2019.

[9] J. Zhang, Y. Ji, S. Jia, H. Li, X. Yu and X. Wang, "Reconfigurable optical mobile fronthaul networks for coordinated multipoint transmission and reception in 5G," *Journal of Optical Communications and Networking*, vol. 9, no. 6, pp. 489-497, June 2017.

[10] J. Zhang, Y. Ji, X. Xu, H. Li, Y. Zhao and J. Zhang, "Energy efficient baseband unit aggregation in cloud radio and optical access networks," *Journal of Optical Communications and Networking*, vol. 8, no. 11, pp. 893-901, Nov. 2016.

[11] D. Rafique and L. Velasco, "Machine learning for network automation: overview, architecture, and applications [Invited Tutorial]," *Journal of Optical Communications and Networking*, vol. 10, no. 10, pp. D126-D143, Oct. 2018.

[12] F. Musumeci et al., "An Overview on Application of Machine Learning Techniques in Optical Networks," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1383-1408, Second quarter 2019.

[13] Optimization vs. heuristics: Which is the right approach for your business? https://icrontech.com/blog_item/optimization-vs-heuristics-which-is-the-right-approach-for-your-business.

[14] N. Carapellese, M. Tornatore and A. Pattavina, "Energy-Efficient Baseband Unit Placement in a Fixed/Mobile Converged WDM Aggregation Network," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 8, pp. 1542-1551, Aug. 2014.

[15] F. Musumeci, C. Bellanzon, N. Carapellese, M. Tornatore, A. Pattavina and S. Gosselin, "Optimal BBU Placement for 5G C-RAN Deployment Over WDM Aggregation Networks," *Journal of Lightwave Technology*, vol. 34, no. 8, pp. 1963-1970, 15 April15, 2016.

[16] R. Riggio, D. Harutyunyan, A. Bradai, S. Kuklinski and T. Ahmed, "SWAN: Base-band units placement over reconfigurable wireless front-hauls," 2016 12th International Conference on Network and Service Management (CNSM), Montreal, QC, pp. 28-36, 2016.

[17] S. S. Lisi, A. Alabbasi, M. Tornatore and C. Cavdar, "Cost-effective migration towards C-RAN with optimal fronthaul design," 2017 IEEE International Conference on Communications (ICC), Paris, 2017.

[18] T. Taleb, A. Ksentini and B. Sericola, "On Service Resilience in Cloud-Native 5G Mobile Systems," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 483-496, March 2016.

[19] M. Y. Lyazidi, L. Giupponi, J. Mangues-Bafalluy, N. Aitsaadi and R. Langar, "A Novel Optimization Framework for C-RAN BBU Selection Based on Resiliency and Price," 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), Toronto, ON, 2017.

[20] B. M. Khorsandi, F. Tonini and C. Raffaelli, "Design methodologies and algorithms for survivable C-RAN," 2018 International Conference on Optical Network Design and Modeling (ONDM), Dublin, 2018.

[21] Shehata, Mohamed, F. Musumeci, and M. Tornatore. "Resilient BBU placement in 5G C-RAN over optical aggregation networks." *Photonic Network Communications*, vol. 37, no. 3, pp. 388-398, June 2019.

[22] H. Hirayama, Y. Tsukamoto, S. Nanba and K. Nishimura, "RAN Slicing in Multi-CU/DU Architecture for 5G Services," 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall), 2019, pp. 1-5, doi: 10.1109/VTCFall.2019.8891584.

[23] Y. Xiao, J. Zhang and Y. Ji, "Energy-efficient DU-CU Deployment and Lightpath Provisioning for Service-oriented 5G Metro Access/Aggregation Networks," in Journal of Lightwave Technology, doi: 10.1109/JLT.2021.3069897.

[24] D. Silver et al., "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484-489, January 2016.

[25] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529-533, 2015.

[26] X. Chen, R. Proietti and S. J. B. Yoo, "Building Autonomic Elastic Optical Networks with Deep Reinforcement Learning," in *IEEE Communications Magazine*, vol. 57, no. 10, pp. 20-26, October 2019, doi: 10.1109/MCOM.001.1900151.

[27] M. R. Raza, C. Natalino, P. Öhlen, L. Wosinska and P. Monti, "Reinforcement Learning for Slicing in a 5G Flexible RAN," in Journal of Lightwave Technology, vol. 37, no. 20, pp. 5161-5169, 15 Oct.15, 2019, doi: 10.1109/JLT.2019.2924345.

[28] X. Zhang, B. Li, J. Peng, X. Pan and Z. Zhu, "You Calculate and I Provision: A DRL-Assisted Service Framework to Realize Distributed and Tenant-Driven Virtual Network Slicing," in Journal of Lightwave Technology, vol. 39, no. 1, pp. 4-16, 1 Jan.1, 2021, doi: 10.1109/JLT.2020.3023693.

[29] Y. Cao, Y. Zhao, J. Li, R. Lin, J. Zhang and J. Chen, "Multi-Tenant Provisioning for Quantum Key Distribution Networks With Heuristics and Reinforcement Learning: A Comparative Study," in IEEE Transactions on Network and Service Management, vol. 17, no. 2, pp. 946-957, June 2020, doi: 10.1109/TNSM.2020.2964003.

[30] Y. Xiao, J. Zhang, Z. Gao, and Y. Ji, "Service-oriented DU-CU Placement Using Reinforcement Learning in 5G/B5G Converged Wireless-Optical Networks," in Optical Fiber Communication Conference (OFC) 2020, OSA Technical Digest (Optical Society of America, 2020), paper T4D.5.

[31] 3GPP TS 38.300 V0.4.1, "NR; NR and NG-RAN Overall Description; Stage 2 (Release 15)," June 2017.

[32] Van Hasselt, Hado, A. Guez, and D. Silver. "Deep Reinforcement Learning with Double Q-learning," Computer Science, 2015.

[33] Yen, Jin Y. "Finding the k shortest loopless paths in a network." management Science 17.11 (1971): 712-716.

[34] R. S. Sutton, A. G. Barto, "Reinforcement learning: An introduction," 2 ed., MIT press, pp. 62–63.

[35] IEEE 1914.1, "Dimensioning Challenges of xhaul", Tech. Rep., V2.0, Mar. 2018.